Heart Disease Risk Prediction System

Prepared by: Muddapati Sai Teja
Role: Machine Learning Intern
Organization: RD INFRO TECHNOLOGY

1. Project Overview

Cardiovascular diseases are one of the leading causes of death worldwide. Early detection of heart disease plays a crucial role in improving patient survival rates and reducing healthcare costs. This project focuses on building an end-to-end **Machine Learning–based Heart Disease Risk Prediction System** that can predict whether a patient is likely to have heart disease based on clinical parameters.

The project was executed as part of an internship at **RD INFRO TECHNOLOGY**, following a structured task-based approach from problem understanding to deployment.

2. Problem Statement

To design and deploy a machine learning model that predicts the risk of heart disease using patient health data such as age, blood pressure, cholesterol levels, heart rate, and other clinically relevant indicators.

3. Objectives

- Understand the medical problem and dataset characteristics
- Collect and validate a reliable healthcare dataset
- Clean and preprocess the data for modeling
- Perform exploratory data analysis (EDA)
- Engineer meaningful features
- Train, evaluate, and tune machine learning models
- Deploy the final model as a user-friendly web application

4. Dataset Description

- **Source:** UCI Heart Disease Dataset (via Kaggle)
- **Records:** ~300 patient records
- **Features:** Demographic, clinical, and test-based attributes

- **Target Variable:** Presence or absence of heart disease

The dataset was chosen due to its medical relevance, credibility, and wide usage in academic and industry projects.

---

5. Project Workflow (Task-wise Summary)

Task 1: Problem Understanding & ML Objective

- Defined the healthcare problem
- Identified supervised classification as the ML approach
- Established business and technical objectives

Task 2: Data Collection & Verification

- Imported dataset using Pandas
- Verified shape, columns, data types
- Checked for missing values and duplicates
- Ensured data readiness for preprocessing

Task 3: Data Cleaning & Preprocessing

- Handled missing values
- Encoded categorical variables
- Scaled numerical features using StandardScaler
- Split data into training and testing sets
- Saved preprocessing artifacts for reuse

Task 4: Exploratory Data Analysis (EDA)

- Analyzed feature distributions
- Studied relationships between features and target variable
- Visualized trends using graphs and plots
- Derived insights useful for feature selection

Task 5: Feature Engineering & Selection

- Created new meaningful features
- Evaluated feature importance
- Selected the most impactful features for modeling
- Reduced dimensionality while preserving performance

Task 6: Model Training & Evaluation

- Trained multiple models (Logistic Regression, Random Forest, SVM, XGBoost)

- Evaluated models using accuracy, precision, recall, F1-score, ROC-AUC
- Compared results and shortlisted the best-performing model

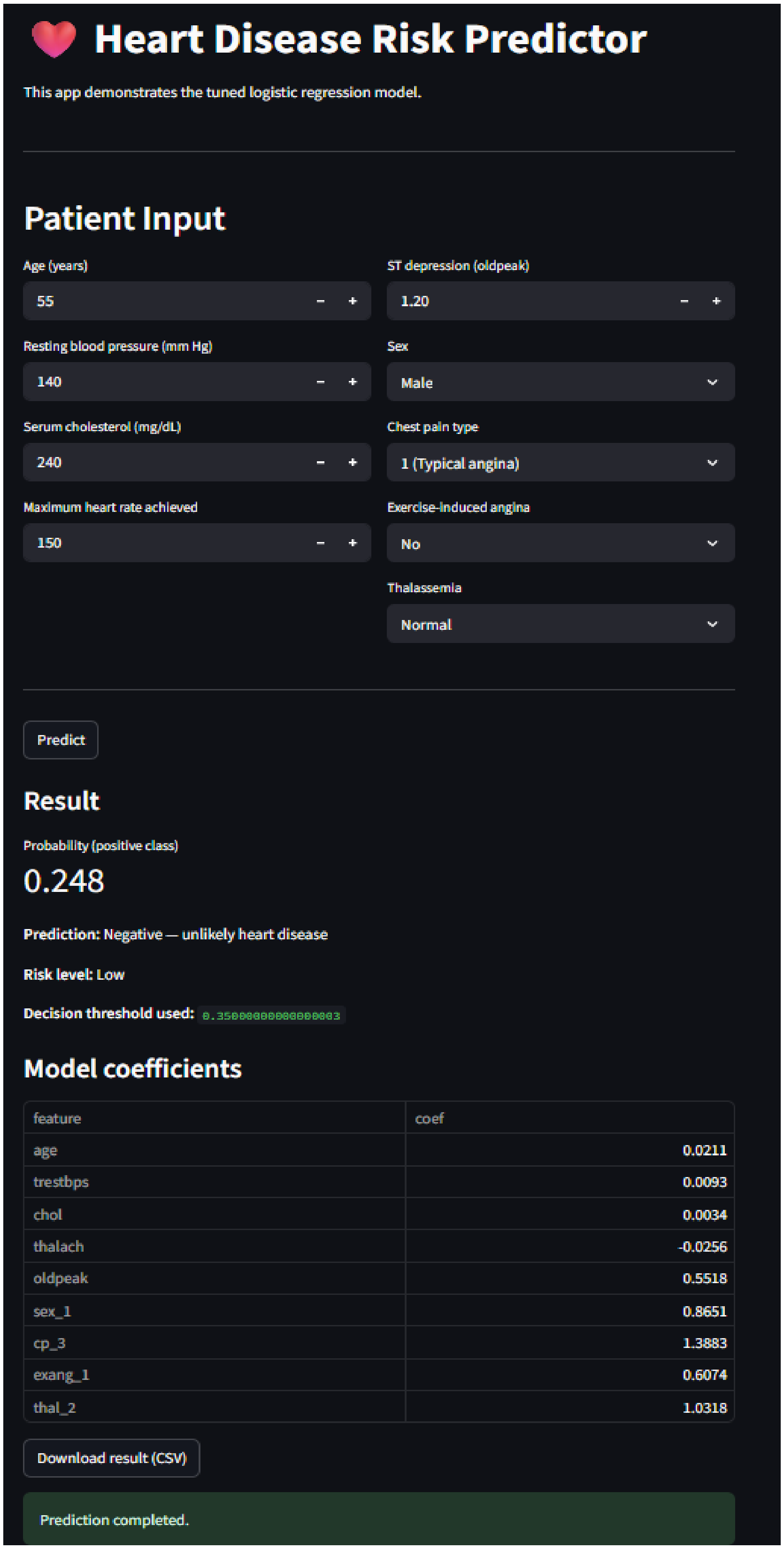
Task 7: Model Tuning & Threshold Optimization

- Tuned the selected model
- Optimized decision threshold based on business needs
- Improved recall without significantly impacting precision
- Saved the final tuned model for deployment

Task 8: Model Deployment

- Built a professional **Streamlit web application**
- Designed user-friendly input fields with medical explanations
- Integrated preprocessing and prediction pipeline
- Deployed the app on **Streamlit Cloud**

Live Application Link:
https://heart-disease-app-app-jcfqugmaexmnr4gpndihzx.streamlit.app/



Heart Disease Risk Predictor

This app demonstrates the tuned logistic regression model.

Patient Input

Age (years)
55

Resting blood pressure (mm Hg)
140

Serum cholesterol (mg/dL)
240

Maximum heart rate achieved
150

ST depression (oldpeak)
1.20

Sex
Male

Chest pain type
1 (Typical angina)

Exercise-induced angina
No

Thalassemia
Normal

Result

Probability (positive class)

0.248

Prediction: Negative — unlikely heart disease

Risk level: Low

Decision threshold used: `0.35000000000000003`

Model coefficients

feature	coef
age	0.0211
trestbps	0.0093
chol	0.0034
thalach	-0.0256
oldpeak	0.5518
sex_1	0.8651
cp_3	1.3883
exang_1	0.6074
thal_2	1.0318

Prediction completed.

6. Final Model Summary

- **Model Used:** Tuned Logistic Regression
- **Why Logistic Regression?**
  - High interpretability (important in healthcare)
  - Strong and stable performance
  - Easy to explain predictions to non-technical users
- **Key Metrics:**
  - High F1-score and ROC-AUC
  - Optimized recall to reduce false negatives

---

7. Application Output

- Probability score for heart disease
- Final prediction (Positive / Negative)
- Risk level classification (Low / Medium / High)
- Transparent display of decision threshold

---

8. Tools & Technologies Used

- **Programming:** Python
- **Libraries:** Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn
- **Model Deployment:** Streamlit
- **Development Environment:** VS Code
- **Version Control:** Git & GitHub

---

9. Conclusion

This project demonstrates the complete lifecycle of a real-world machine learning application in the healthcare domain. From data collection to deployment, each stage was carefully designed and executed following industry best practices.

The deployed application successfully predicts heart disease risk and can assist in early screening and decision support.

---

10. Future Enhancements

- Integration with real-time hospital data

- Use of advanced ensemble models
- Model explainability using SHAP or LIME
- Deployment on cloud platforms with authentication
- Mobile-friendly UI

Declaration:
This project was independently developed by *Muddapati Sai Teja* as part of a Machine Learning internship at **RD INFRO TECHNOLOGY**.